Optimal γ and C for ϵ-Support Vector Regression with RBF Kernels

Longfei Lu*

June 15, 2015

*Longfei.Lu@live.com

Abstract

The objective of this study is to investigate the efficient determination of C and γ for Support Vector Regression with RBF or mahalanobis kernel based on numerical and statistician considerations, which indicates the connection between C and kernels and demonstrates that the deviation of geometric distance of neighbour observation in mapped space effects the predict accuracy of ϵ-SVR. We determinate the arrange of γ & C and propose our method to choose their best values.

Introduction

Traditional forecasting algorithm like ARIMA, Exponential Smoothing can provide good forecasting results with regard to trend, season and other linear correlated features . In practice, those features are normally non-linear. To solve non-linear forecasting problems, ϵ-Support vector regression (ϵ-SVR) is employed. Support vector machine (SVM) is nowadays wildly used for classification problems in many areas. However, ϵ-SVR is hardly used because of the uncertain parameter C, ϵ for its dual problem. By using RBF or Mahalanobis kernels, value of γ decides determination of kernel matrix and hence is the key of whole system. An overview of choosing those parameters is given by [1]. Best selection of C is given by [6], which can be done with in limited iterations. The most used methods are searching methods like random search, grid search, pattern search. Cherkassy and Ma (2004)[2] have proposed one way to determinate those parameters directly from the data. But there is still one parameter c need to be extra searched within pre-existing arrange. Our propose is also to determinate C and γ directly but without using any searching methods. In this paper, we give a short overview of ϵ-SVR in section 1 and discuss the determination of C and γ in section 2 and 3. Then we test our algorithms using practice data in section 4 and summery results in section 5.

1 ϵ-Support Vector Regression

Assume $(x_1, y_1), (x_2, y_2), ..., (x_N, y_N), \ldots \in \mathcal{X} \times \mathbf{R} \subseteq \mathbf{R}^{n+1}$ are observation pairs, $x_i \in \mathcal{X} \subset \mathbf{R}^n$ is feature vector and $y_i \in \mathbf{R}$ is the target output. Define $N_{\mathcal{X}}$ as the total number of all observation in training set. According to [3]

the dual problem form of ϵ-SVR under given C with kernel K is

$$\begin{aligned} \textbf{min} \quad & \frac{1}{2}(\underline{\alpha}-\underline{\alpha}^*)^T Q(\underline{\alpha}-\underline{\alpha}^*)+\epsilon\sum(\alpha_i+\alpha_i^*)+\sum y_i(\alpha_i-\alpha_i^*) \\ \textbf{subject to} \quad & \underline{e}^T(\underline{\alpha}-\underline{\alpha}^*)=0, \\ & 0\le\alpha_i,\alpha_i^*\le C, i=1,...,N_s, \end{aligned} \tag{1}$$

where $Q_{i,j} := K(x_i, x_j)$ and $\underline{\alpha} = (\alpha_1, ..., \alpha_{N_s})$, $\underline{\alpha}^* = (\alpha_1^*, ..., \alpha_{N_s}^*)$. The corresponding approximate function is

$$f(x) \;=\; \sum_{i=1}^{N_s}(-\alpha_i+\alpha_i^*)K(x_i,x)+b, \tag{2}$$

where N_s is the total number of Support Vectors which are obtained from input parameter ϵ. We consider ϵ as accuracy indicator or acceptable tolerance within this system. In practice, we use natural defined tolerance as input value for ϵ.

2 Optimal γ

A feature map $\phi: x \mapsto \phi(x) \in \mathbf{R}^m, m \ge n$ builds new norm with respect to kernel K:

$$\begin{aligned} \|\phi(x_i)-\phi(x_j)\|^2 &= K(x_i,x_i)+K(x_j,x_j)-2K(x_i,x_j) \\ &= 2-2K(x_i,x_j), \text{ for RBF kernel} \\ & \forall x_i, x_j \in \widehat{X} \subset \mathbf{R}^n. \end{aligned}$$

In mapped space, the splitting of two neighbour observation , which means the deviation of mapped features should be large. Small values of deviation mean mapped features have very few difference and lead to linear dependences of kernel matrix in practice which produce few independent features and enlarge the solutions space. it is not difficult to show that features in RBF Kernel mapped features have same structure as in original which means

$$\|x_i-x_j\| \le \|x_k-x_l\| \Leftrightarrow \|\phi(x_i)-\phi(x_j)\| \le \|\phi(x_k)-\phi(x_l)\|.$$

More independence of mapped features lead to small solution spaces. One measure for this character is the deviation of all 2-te Norms between every

two observation. Define **Deviation Function** of observation as

$$\boldsymbol{L(\gamma)} := \frac{1}{N_{\mathcal{X}}(N_{\mathcal{X}}-1)} \sum_{(i,j)\in\mathcal{X}} \left(\|\phi(x_i) - \phi(x_j)\| - \frac{1}{N_{\mathcal{X}}(N_{\mathcal{X}}-1)} \sum_{(k,l)\in\mathcal{X}} \|\phi(x_k) - \phi(x_l)\| \right)^2.$$

The Deviation Function $L(\gamma)$ describes geographic differences of whole observation in mapped space with help of 2-te Norm and is convex. Our best choice of γ_{opt} is the point when the $L(\gamma)$ reaches its maximum where $L^{'}(\gamma_{opt}) = 0$ with fist derivation formal

$$\begin{aligned} L^{'}(\gamma) &= \frac{2}{N_{\mathcal{X}}(N_{\mathcal{X}}-1)} \sum_{(i,j)\in\mathcal{X}} \left(\widehat{L}(x_i, x_j) - \frac{1}{N_{\mathcal{X}}(N_{\mathcal{X}}-1)} \sum_{(k,l)} \widehat{L}(x_k, x_l) \right) \\ &\quad \cdot \left(\widehat{D}(x_i, x_j) - \frac{1}{N_{\mathcal{X}}(N_{\mathcal{X}}-1)} \sum_{(k,l)\in\mathcal{X}} \widehat{D}(x_k, x_l) \right), \end{aligned}$$

and its second derivation formal is

$$\begin{aligned} L''(\gamma) &= \frac{2}{N_{\mathcal{X}}(N_{\mathcal{X}}-1)} \sum_{(i,j)\in\mathcal{X}} \left(\widehat{L}'(x_i, x_j) - \frac{1}{N_{\mathcal{X}}(N_{\mathcal{X}}-1)} \sum_{(k,l)\in\mathcal{X}} \widehat{L}'(x_k, x_l) \right) \\ &\quad \cdot \left(\widehat{D}(x_i, x_j) - \frac{1}{N_{\mathcal{X}}(N_{\mathcal{X}}-1)} \sum_{(k,l)\in\mathcal{X}} \widehat{D}(x_k, x_l) \right) \\ &+ \frac{2}{N_{\mathcal{X}}(N_{\mathcal{X}}-1)} \sum_{(i,j)\in\mathcal{X}} \left(\widehat{L}(x_i, x_j) - \frac{1}{N_{\mathcal{X}}(N_{\mathcal{X}}-1)} \sum_{(k,l)\in\mathcal{X}} \widehat{L}(x_k, x_l) \right) \\ &\quad \cdot \left(\widehat{D}'(x_i, x_j) - \frac{1}{N_{\mathcal{X}}(N_{\mathcal{X}}-1)} \sum_{(k,l)\in\mathcal{X}} \widehat{D}'(x_k, x_l) \right) \\ &= \frac{2}{N_{\mathcal{X}}(N_{\mathcal{X}}-1)} \sum_{(i,j)\in\mathcal{X}} \left(\widehat{D}(x_i, x_j) - \frac{1}{N_{\mathcal{X}}(N_{\mathcal{X}}-1)} \sum_{(k,l)\in\mathcal{X}} \widehat{D}(x_k, x_l) \right)^2 \\ &+ \frac{2}{N_{\mathcal{X}}(N_{\mathcal{X}}-1)} \sum_{(i,j)\in\mathcal{X}} \left(\widehat{L}(x_i, x_j) - \frac{1}{N_{\mathcal{X}}(N_{\mathcal{X}}-1)} \sum_{(k,l)\in\mathcal{X}} \widehat{L}(x_k, x_l) \right) \\ &\quad \cdot \left(\widehat{D}'(x_i, x_j) - \frac{1}{N_{\mathcal{X}}(N_{\mathcal{X}}-1)} \sum_{(k,l)\in\mathcal{X}} \widehat{D}'(x_k, x_l) \right), \end{aligned}$$

with respect to

$$\widehat{L}(x_i, x_j) = \left(2 - 2\exp(-\gamma G(x_i, x_j)) \right)^{\frac{1}{2}}$$

$$\widehat{D}(x_i, x_j) = G(x_i, x_j) \exp(-\gamma G(x_i, x_j))(2 - 2 \cdot \exp(-\gamma G(x_i, x_j))^{-\frac{1}{2}}.$$

$$\begin{aligned}
\widehat{D}'(x_i, x_j) &= -(G(x_i, x_j))^2 \exp(-\gamma G(x_i, x_j)) \left(2 - 2\exp(-\gamma G(x_i, x_j))\right)^{-\frac{1}{2}} \\
&\quad \cdot \left(1 + \exp(-\gamma G(x_i, x_j))(2 - 2\exp(-\gamma G(x_i, x_j)))^{-1}\right) \\
&= -G(x_i, x_j)\widehat{D}(x_i, x_j) \left(1 + \exp(-\gamma G(x_i, x_j))(\widehat{L}(x_i, x_j))^{-2}\right)
\end{aligned}$$

For RBF Kernel $K(x_i, x) = \exp(-\gamma \|x_i - x\|^2)$:

$$G(x_i, x_j) = \|x_i - x_j\|^2.$$

For Mahalanobis Kernel $K(x_i, x) = \exp(-\gamma \frac{1}{m}(x_i - x)Q^{-1}(x_i - x))$:

$$G(x_i, x_j) = \frac{1}{m}(x_i - x_{i+1})Q^{-1}(x_i - x_j),$$

with

$$\begin{aligned}
Q &= \frac{1}{N_{\mathcal{X}}} \sum_{k=1}^{N_{\mathcal{X}}} (x_k - c)(x_k - c)^T, \\
c &= \frac{1}{N_{\mathcal{X}}} \sum_{k=1}^{N_{\mathcal{X}}} x_k, \\
m &= \frac{1}{N_{\mathcal{X}}} \sum_{k=1}^{N_{\mathcal{X}}} (x_k - c)^T Q^{-1}(x_k - c).
\end{aligned}$$

Because of the convexity of deviation function, its maximum always exists and by using numeric method like Newton Method we obtain very good balance between computation performance and model accuracy. Our proposed deviation function $L(\gamma)$ is also very helpful for determining arrange for searching method (see figure 1). The arrange of γ is $(0, \gamma_h)$, where $\gamma_h = \arg\min L(\gamma)$. At point γ_h, the value of kernel function $K(x_i, x_j) \approx 0$, for $i \neq j$, which leads to strong over-fitting problem (see figure 6). The best value of γ is at $\arg\max L(\gamma)$.

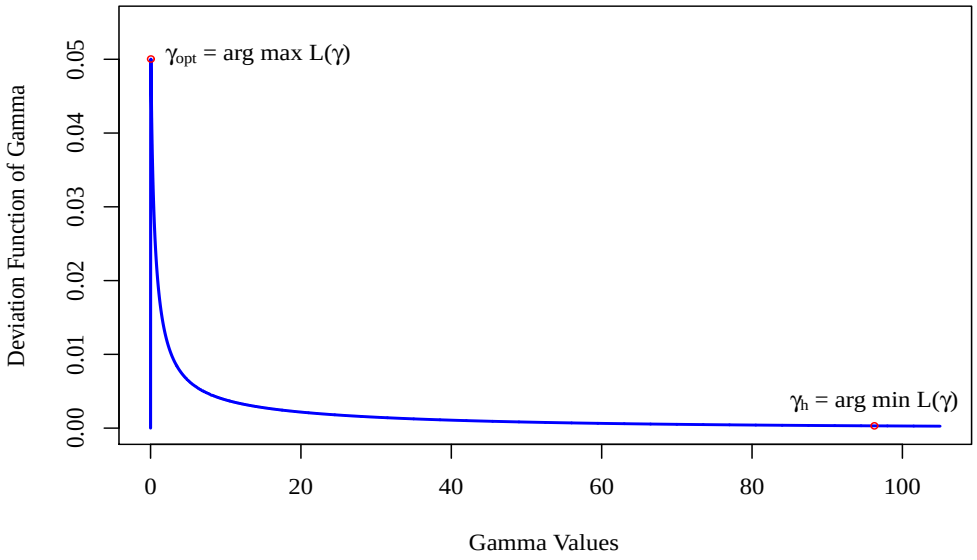


Figure 1: Deviation function $L(\gamma)$ for RBF kernel

3 Optimal choice of C

We use the proposed method from [6] to get best value of C, which provides very stable solution. In order to reduce the total number of iterations, we give a reasonable initial C before it begins by applying mean value theorem in mapped space. From (2) exists $x \in \mathcal{X}$ satisfying

$$\|y_i - y_j\| = \|f'(\phi(x)) \cdot (\phi(x_i) - \phi(x_j))\|, \forall i, j \in \{1, .., N_{\mathcal{X}}\}.$$

For RBF Kernel, we can proof that

$$\frac{\|y_i - y_j\|}{\|\phi(x_i) - \phi(x_j)\|} \leq \|f'(\phi(x))\| \leq N_s C. \tag{3}$$

For worst situation that $N_s = 1$ and $\|\phi(x_i) - \phi(x_j)\| = \exp(-\gamma\|x_i - x_j\|^2)$, we obtain

$$C \geq \max_{i,j \in \{1,..,N_{\mathcal{X}}\}} \|y_i - y_j\| \exp(\gamma\|x_i - x_j\|^2). \tag{4}$$

We use right side of (4) as initial C_{ini} for iteration of finding best C. New value of C after every solving of SVR is calculated by

$$C_{new} = \frac{N_{\mathcal{X}} + N_s}{\sum\limits_{x_i \in X_C} L_\epsilon(y_i - f(x_i)) + \sum\limits_{x_i \in X_M} \frac{1}{C - |\alpha_i - \alpha_i^*|} + \frac{\epsilon N_{\mathcal{X}}}{\epsilon C + 1}} \tag{5}$$

with

$$\begin{aligned} X_C &= \{x_i |\ |y_i - f(x_i)| > \epsilon \text{ with } \alpha_i = C, \alpha_i^* = 0 \text{ or } \alpha_i = 0, \alpha_i^* = C\}, \\ X_M &= \{x_i |\ |y_i - f(x_i)| = \epsilon \text{ with } 0 < \alpha_i < C, \text{ or } 0 < \alpha_i^* < C\} \end{aligned}$$

and ϵ loss function L_ϵ is defined as

$$L_\epsilon(y_i - f(x_i)) = \begin{cases} 0 & \text{for } |y_i - f(x_i)| \leq \epsilon \\ |y_i - f(x_i)| - \epsilon & \text{otherwise.} \end{cases} \tag{6}$$

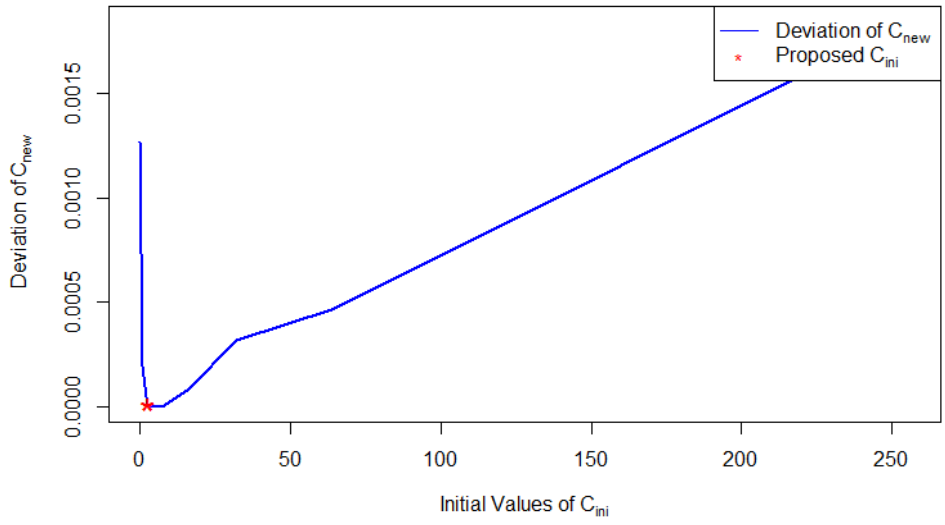


Figure 2: Devation of C_{new} with respect to different initial value C_{ini}

According to our experiments, our proposed initial value has relative small deviation during computation of C_{new} (see figure 2), which reduces the

iteration number by setting stop critical as Δ changes between new and last C_{new}. Experiments also show that no matter how big C_{ini} is, it archives its stable value within limit iterations. Figure 3 and Table 1 show that C value changes not so much just after second solving SVR in our experiments by input data of gas consumption with temperature and weekdays as features.

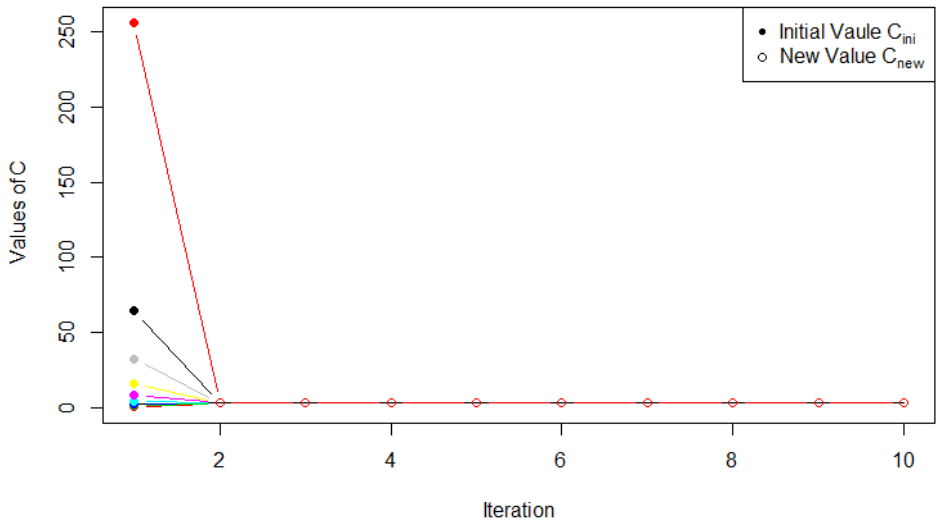


Figure 3: Iteration of finding stable C with different C_{ini}

C_{ini}	1st C_{new}	2nd C_{new}	3rd C_{new}	4th C_{new}	5th C_{new}	6th C_{new}	7th C_{new}	8th C_{new}	9th C_{new}	10th C_{new}
2.582123	2.887401	2.895912	2.893142	2.903892	2.898490	2.898554	2.898552	2.898462	2.898543	2.893128
0.100000	2.738660	2.897665	2.898563	2.898521	2.898552	2.898515	2.898516	2.898517	2.898550	2.893649
1.000000	2.836104	2.893117	2.893175	2.903887	2.898516	2.898489	2.898544	2.898492	2.898560	2.898525
2.000000	2.866985	2.898480	2.898493	2.898526	2.898519	2.898527	2.898528	2.898566	2.898478	2.898450
4.000000	2.891488	2.898533	2.898536	2.898512	2.898500	2.898469	2.898476	2.898543	2.898502	2.898529
8.000000	2.892421	2.893184	2.898528	2.898532	2.898523	2.898495	2.898494	2.898525	2.898560	2.898519
16.000000	2.936898	2.893208	2.893217	2.893139	2.903910	2.898495	2.898543	2.898531	2.898510	2.898463
32.000000	2.977581	2.897040	2.898510	2.898527	2.898548	2.897689	2.898522	2.898533	2.898511	2.898530
64.000000	2.995011	2.905414	2.902729	2.898555	2.898485	2.898533	2.898483	2.898527	2.898541	2.898588
256.000000	3.090617	2.909661	2.898573	2.898489	2.898521	2.898504	2.898506	2.898570	2.898555	2.898496

Table 1: Iterations Table

4 Experiments

All experiments were performed by using practice data from an energy company from 2009-01-01 to 2011-12-31, which contains 15 different features. The training set was set from 2009-01-01 to 2011-09-24. We determinated the curves of γ according to the description in section 3 and used γ at maximum of $L(\gamma)$ for determination of parameter C. Back test was performed by using data from 2011-09-25 to 2011-12-31 for RBF kernels. Those results were compared with results generated by searching $\gamma \in \{2^{-15}, ..., 2^3\}$ and $C \in \{2^{-5}, ..., 2^{15}\}$. Package "e1071" of R was used. We scaled arrange of data into $[0, 1]$ and used default scale option for further calculation. We employed root-mean-square error (RMSE)

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N}(y_i - f(x_i))^2}{N}}$$

and mean absolute percentage error (MAPE)

$$\text{MAPE} = \frac{100\%}{N}\sum_{i=1}^{N}\Big|\frac{y_i - f(x_i)}{y_i}\Big|.$$

as measures of accuracy of ϵ-SVR. Except those two famous measures, we define a total new measure **System Error Deviation Measure** (SEDM) as

$$\text{SEDM} = \frac{1}{N}\sum_{i=1}^{N}\Big((y_i - f(x_i)) - \frac{1}{N}\sum_{j=1}^{N}(y_j - f(x_j))\Big). \tag{7}$$

			Training		**Back Test**	
	γ	C	RMSE	MAPE	RMSE	MAPE
Our Propose:	0.03224	0.42112	0.03808	13.5430%	0.05082	15.4615%
Tune Method:	0.04000	0.99000	0.03664	12.6075%	0.05262	15.7146%
Best Solution:	0.02	0.61	0.03758	13.6883%	0.05005	14.9604%
Over-fitting:	96.27366	55.30339	5.06542e-05	0.0279%	0.16210	55.3389%

Table 2: Results with $\epsilon = 0$ for figure 4

Table 2 shows that our propose produced better results than tune method, which has solved SVR over 10,000 times and has the smallest value in training area. The γ value of best solution is smaller than our propose, which always has the smallest RMSE value in back test area. Setting of over-fitting is $\gamma = \arg\min L(\gamma)$. Figure 4 shows the whole γ-C-RMSE space. Our propose locates in the same level as grid search and best solution.

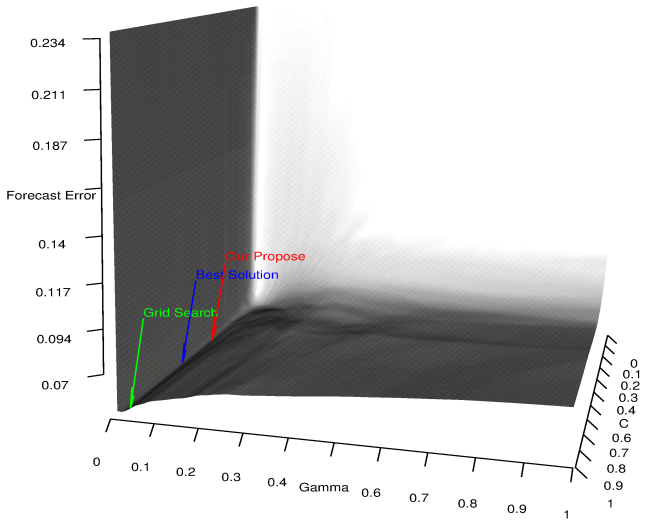
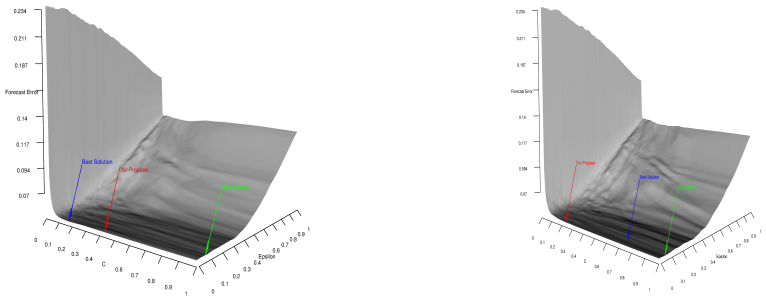


Figure 4: Solution Space $\epsilon = 0$ for Table 2

Table 3 shows the case that $\epsilon \neq 0$. We tuned all three parameter extra within $[10^{-10}, 1] \times [10^{-10}, 1] \times [0, 1]$ for $\gamma, C,$ and ϵ, which has solved SVR over 1,000,000 times. We chose smallest RMSE value in training for Extra Tuning. But its forecasting error is worse than results in Table 2. Therefore, we used the γ_{best} from Best Solution, γ_{opt} from Our Propose in Table 2 and each was tuned within interval $[10^{-10}, 1]$ for C and $[0, 0.99]$ for ϵ. Our propose for γ_{opt} produced better results in Table 3 with $\epsilon = 0.00516$ than Tune Method, which produced better results for γ_{best}. Figure 4 shows locations of our experimented solutions in C-ϵ-RMSE space for each γ_{opt} and γ_{best}.

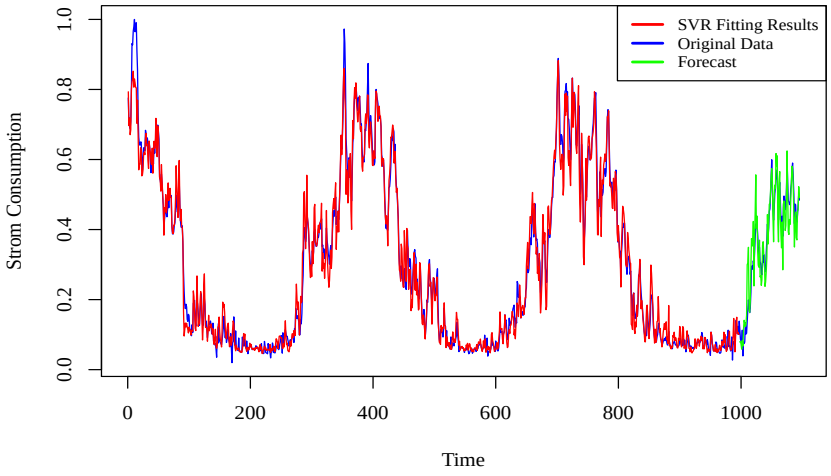
				Training		Back Test	
	γ	C	ϵ	RMSE	MAPE	RMSE	MAPE
Our Propose							
γ_{opt}	0.03224	0.42112	0.00516	0.03808	13.5563%	0.05079	15.4560%
γ_{best}	0.02	0.22035	0.02768	0.03967	14.9641%	0.05038	15.6418%
Tune Method							
γ_{opt}	0.03224	0.99	0.07	0.03635	14.0366%	0.05185	15.9645%
γ_{best}	0.02	0.99	0.05	0.03681	13.9567%	0.05013	14.7957%
Best Solution							
γ_{opt}	0.03224	0.16	1e-10	0.04058	14.4724%	0.05026	15.5602%
γ_{best}	0.02	0.71	0.05	0.03725	14.1742%	0.04992	14.8701%
Extra tuning	0.06	0.99	0.09	0.03599	14.3002%	0.05584	18.0309%

Table 3: Results with optimal ϵ for figure 5

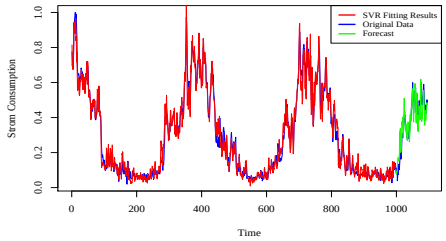


(a) γ_{opt} (b) γ_{best}

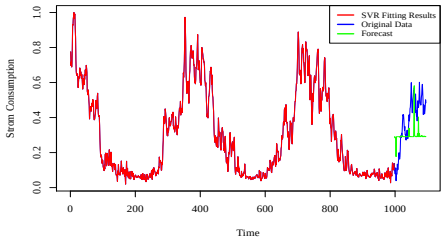
Figure 5: Solution Space for Table 3



(a) $\gamma = \arg\max L(\gamma)$



(b) Using searching method



(c) $\gamma = \arg\min L(\gamma)$ (Over-fitting Problem)

Figure 6: Results generated by package "e1071" in R

5 Conclusion

Deviation function $\boldsymbol{L}(\gamma)$ and set $\boldsymbol{C}$ give the arrange of γ and C, also provide possible solutions to choose their optimal values. By searching method, RMSE of training set became smaller than before.

Reference

[1] Marcos Marin-Galiano Stefan Roeuping Andreas Christmann, Karsten Luebke. Determination of hyper-parameters for kernel based classification and regression.

[2] Vladimir Cherkassky and Yunqian Ma. Practical selection of svm parameters and noise estimation for svm regression.

[3] Chih-Jen Lin Chih-Chung Chang. Libsvm: A library for support vector machines. April 4, 2012.

[4] Chih-Chung Chang Chih-Wei Hsu and Chih-Jen Lin. A practical guide to support vector classification. 2010.

[5] Chih jen Lin and Ruby C. Weng. Simple probabilistic predictions for support vector regression. 2004.

[6] Chris J. Harris Junbin B. Gao, Steve R. Gunn and Martin Brown. A probabilistic framework for svm regression and error bar estimation. 2001.